Homework 1

SHASHANK SHEKHAR – sshekha4
PURVA VASUDEO – ppvasude
RAHUL AETTAPU - raettap

1) Data Properties
   a) Attribute Classification:
      i) Blood Group: Nominal [Discrete]. The blood groups are distinct nominal values such as A+ve,B+ve etc.
      ii) Ticket Number for Raffle Draws: Nominal [Discrete]. Again, the ticket numbers are distinct discrete values which in a sense have no inherent ordering as it depends on the draw. Eg. 789,665,334.
      iii) Brightness: Ratio[Continuous]. Brightness has a meaningful 0 value to signify no light. Hence it is a ratio attribute. Also there can be infinite values of brightness spread across infinite rational values. Eg. 3.56776576 lux
      iv) Grades: Nominal [Binary]. Grades are nominal and in terms of just Pass/Fail they are binary also since they can take only two possible values. There is no measure of difference between them.
      v) Time zones: Nominal [Discrete]. The time zones are nominal values as it can take only 3 discrete values here EST,PST and CST. We do not know the ordering between them and one is not greater than the other in any way.
      vi) Income earned: Ratio [Continuous]. Income is a ratio attribute as it has a meaningful zero, there can be zero income. It is also continuous as a person's income is finally a rational value and in a sense this attribute can take up infinite possible values. Eg. $567993.8867456.
      vii) Vehicle License plate number: Nominal [Discrete]. This is a nominal value as it has distinctiveness and there is no inbuilt ordering amongst the different values which can be assigned to the attribute. Eg. MH01ND345.
      viii) Distance: Ratio [Continuous]. Distance has a meaningful zero and 'x' distance is half of '2x' distance. It can take up infinite possible rational positive values. Eg. 2345 miles.
      ix) Dorm Room Number: Nominal [Discrete]. This is a nominal attribute as it has distinctness and each dorm number is unique from the others, plus they are distinct integer or string values and are discrete. Eg. A135 room number.
      x) Kelvin temperature: Ratio [Continuous]. This is a ratio attribute as it has a meaningful zero. 'x' kelvin temperature is exactly half of '2x' kelvin temperature. Temperature can take up a range of rational values which are infinite and continue. Eg. 0K

   b) Statistics / Operations on attributes:
      i) Make: [Nominal] -> Mode
      ii) Fuel Type: [Nominal] -> Mode
      iii) # of doors: [Ratio] -> Mean, median,mode,Z-scale Normalization, Binary discretization
      iv) Height: [Ratio] -> Mean, median,mode,Z-scale Normalization, Binary discretization
      v) # of cylinders: [Ratio] -> Mean, median,mode,Z-scale Normalization, Binary discretization

vi) Price: [Ratio] -> Mean, median,mode,Z-scale Normalization, Binary discretization

c) Considering as Ordinal Attribute: If we assume that the total marks for the quiz was 100 and that the marks were scaled down between 0 to 5 based on the student's performance and the difficulty level of the question that the student attempted. Then in that case, we are binning the marks of the students in separate categories. For example, Students obtaining marks between 90 to 100 are given 5 marks, 70 to 90 are given 4 marks, 40 to 70 are given 3 marks, 20 to 40 are given 2 marks and below 20 are given 1 mark.

Considering as Ratio Attribute: If we assume that the total marks are 5 and students score between 0 to 5, then in that case each student gets his absolute score. The scores have interval and an absolute zero. For example, 0 marks is an absolute zero. A student scoring 4 marks gets double the marks as compared to the student scoring 2 marks. Also, there is an equal interval between students A, B and C scoring 2,3 and 4 marks respectively.

2) Data Transformation and Data Quality

a) Data in tabular format

ID	Patient	Treatment	SBP
1	1	A	160
2	1	B	300
3	2	A	120
4	2	B	100
5	3	A	130
6	3	B	NA
7	4	A	NA
8	4	B	130
9	5	A	120
10	5	B	110
11	6	A	NA
12	6	B	100
13	7	A	240
14	7	B	120
15	8	A	140
16	8	B	90

b) Handling Missing Data

i) Strategy 1: Remove patients with any missing values

ID	Patient	Treatment	SBP
1	1	A	160
2	1	B	300
3	2	A	120
4	2	B	100
5	3	A	130
8	4	B	130
9	5	A	120
10	5	B	110

12	6	B	100
13	7	A	240
14	7	B	120
15	8	A	140
16	8	B	90

Advantage: Least computation, as this basically removes a few records to finally reduce the number of records we will be considering.

Disadvantage: We lose on data points. For example, we might want to perform some other sort of computation where this attribute might not be as meaningful. But since we have removed that row, we have lost on a data point.

ii) Strategy 2:

ID	Patient	Treatment	SBP
1	1	A	160
2	1	B	300
3	2	A	120
4	2	B	100
5	3	A	130
6	3	B	143
7	4	A	143
8	4	B	130
9	5	A	120
10	5	B	110
11	6	A	143
12	6	B	100
13	7	A	240
14	7	B	120
15	8	A	140
16	8	B	90

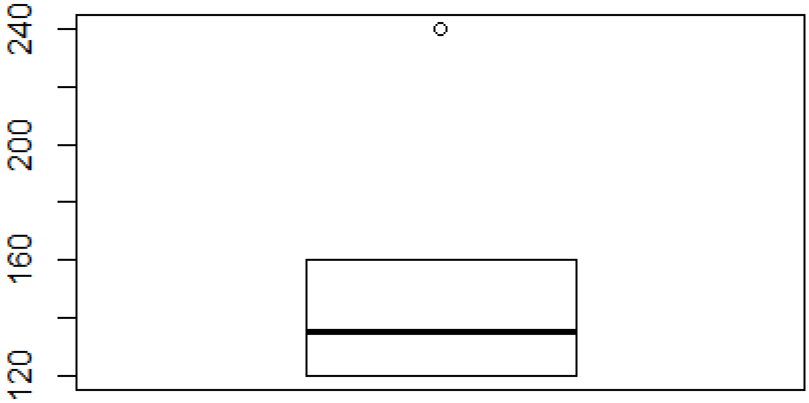
Advantage: We do not lose on any data, assuming that the data should have been there but due to some human error it is not there.

Disadvantage: This method assumes that the data is Missing Completely at Random. Basically, the assumption here is that the value should have existed but due to whatever reason it is not existing. If the data that is missing is not missing at random, then in that case the imputed values may be biased based on which part of the data is missing. For example, a patient might not have gone through Treatment B, he has a record of N/A which actually means that the data should not be there as the patient never really went through that treatment. Hence in this case substituting the average of the existing data points would be a wrong approach.

c) Results of Medical Experiments

i) Considering results that are beyond (Mean +/- 2 SD) away are considered outliers. Values > 264 and values < 22 are considered outliers. Hence, the value of 240 is not an outlier whereas the value of 300 is an outlier. Below are the box plots for the two treatments :

Treatment A :

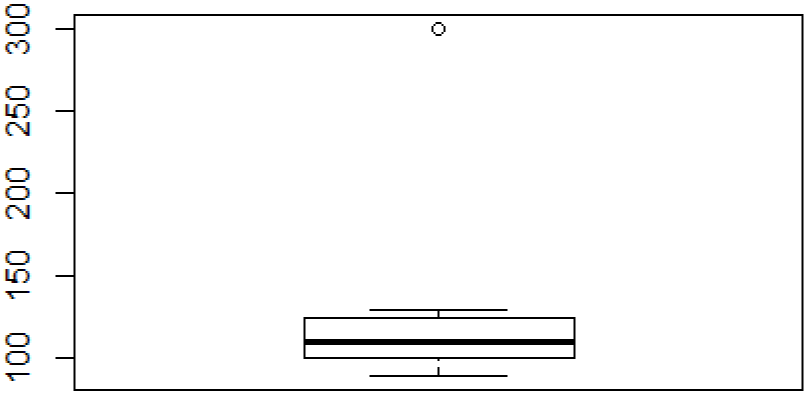


Code :

```
> a_treatment <- c(160, 120, 130, 120, 240, 140)
> boxplot(a_treatment)
```

As can be seen, this box plot has no fences or whiskers. This means the largest value i.e. 240 is the equal to the upper quartile, hence it does not display any such signs of being an outlier. We can't really say if this is noise because noise are random inaccurate values added to the data, and this might not be one.

Treatment B:



Code :

```
> b_treatment <- c(300, 100, 130, 110, 100, 120, 90)
> boxplot(b_treatment)
```

As can be seen the point 300 lies way above the upper quartile or the upper whisker of the box plot. Hence in this case we can clearly say this is an outlier.

Dealing with Outliers: One can try to find out if the outlier value is by mistake or it is actually a very large or small value. If the outlier is by mistake, then in that case the value can be imputed by different methods such as taking mean or by regression. If the value is not due to mistake, then the entire record can be dropped, or the outlier value can be capped to a maximum value.

ii) I believe that the inaccuracies in the reading may create outliers or remove the reading from being an outlier if it was already an outlier but inaccuracies in reading will never create noise in the data.

3) Sampling

a) Stating the sampling method

i) Stratified Sampling: Since the professors are divided into groups and equal number of professors were selected from each group for the study.

ii) Simple Random Sample with replacement: This is because in the final sample there is a value of 2 which is repeated 3 times whereas in the actual population it was present only 2 times.

iii) Adaptive sampling: To reach the accuracy of 90%, we increase the sample size until it meets the required criteria.

b) Senate and House surveys

i) Stratified sampling makes sense over here because the states in a country logically represent groups of people. Since we already have natural grouping in this case(the states), it makes sense to use the stratified sampling method. Moreover, even the Senate and the House of Representatives are considering the states as a group of people out of which a representative is chosen in proportion to the size of the state or the same number of representatives from each state. This is basically stratified sampling itself with different techniques to select from the different groups.

ii) For the Senate, since there is equal representation irrespective of the size of the state, we would put the same logic for selecting the number of participants contributed by Alaska irrespective of the population. As the senate has total of 100 members with 2 from each state, we can decipher that there are 50 states. Each of these 50 states will get an equal number of participants. Hence Alaska will contribute 20 participants out of the 1000 participants.

iii) For the House, there is representation of each state which is proportional to the state's population. From a house of 435 members, Florida has 27 members. This gives us the weightage of Florida in the total house members as 27/435 which is roughly 0.062. Hence we keep the same weightage when deciding on the total number of participants which will be taken from Florida. So in 1000 participants for the House, Florida will contribute 6 participants.

iv) Advantages of the Senate approach to stratified sampling is it is easy to conduct, as from each group we need to pick out the same number of sample data points.

Advantages of the House approach might be that it keeps in mind the corresponding weightage for each group with regards to total number of data samples needed. In this way, there is no unnecessary over representation of a particular group even if its size is really small. Eg. In this case the number of participants offered by Alaska will be proportional to its members or weight in the House of Representatives. In the Senate approach we can say Alaska will be over represented since it gets equal standings with the other states although its population might be half of the other states.

4) Dimensionality Reduction

a) Looking at the graph PC1 seems like a good choice of principal component to retain. This is because, on the graph, this PC is part of a steep slope. In addition to this, this PC has a high eigenvalue which tells us that the variance explained by PC1 is very high. This will mean that this 1st PC alone will explain most of the variance in the data set, given the steep slope of the line. As PC1 has an EigenValue of more than 1, hence this one will explain the most proportion of the variance. PC2 has a higher Eigen Value than PC3 and PC4,but still this value is almost 0, which means this PC does not explain much of the variance in the data.

b) According to PC1, the features that explain the most variance are petal width and petal length. This is because, as given in the table of figure 1, petal width has a weight of 0.855 and petal length has a weight of 0.505 in PC1, which is far greater than the absolute weight of other features in the PC.

c) We would want to retain the PC which are on a steep slope in the graph. This is because these PCs will explain the maximum proportion of variance in the data set. Hence in this case we would select PC1 and PC2 as these two are on the steep slope of the graph, and they will account for more than 90% of the variance as both these PCs have a high eigenvalue which tells us they explain most of the variance in the data set.

d) According to PC1 in figure 2, petal width, petal length and sepal length are the features which contribute most to the variance. This is because all three of these features have an absolute weight of 0.5 around in PC1 which is greater than the contribution of sepal width(absolute weight of 0.3).

e) When we do PCA with raw data, we observe that only PC1 explains most of the variance. While in doing PCA on normalized data, we observe that we have PC1 and PC2 contributing to more than 60% of the variance. I would prefer to do PCA on normalized data, as this takes care of the fact that the variables may not have the same unit of measurement (Example: one variable might take values in Hundreds and the other might take values in Thousands). Normalization will take care of this. If normalization is not done in this case, it might bias our PCA towards a given feature.

f) Based on the results, I would first select the petal width as a feature. This is because in both PCA1 and PCA2, petal width has high weightage in high eigenvalue PCs, which are on steep slopes in the graph as explained earlier. I would also like to select petal length and sepal length, as in PCA2 they also contribute fairly in the PC1 computation, and in this, PC1 has the highest explanation of the variance as compared to the other PCs in PCA2.

5) Discretization

a) In Binning by Equal-width, the range of each interval is constant. The width of the interval is calculated by dividing the range with the bin or partition size. If A and B are the lowest and highest values of the data then the width W is calculated as W = (B-A)/N where N is the number of bins.

In our case we are taking up the TEMPERATURE attribute which has the following data.

TEMPERATURE = [85,80,83,70,68,65,64,72,69,75,75,73,81,71,95,50].

The lowest (A) and highest (B) values are 50 and 95 respectively. Number of bins is 4. Therefore Width W=(95-50)/4. We get width to be W =11.25.

The first interval starts from 50-ε (ε is a small value) as 50 has to be included in the bin and has upper bound of (50 + 11.25) which is 61.25. The first interval is (50-ε,61.25]. The second interval starts from 61.25 and ends at 72.5 (61.25+11.25]. Similarly third interval starts from 72.5 and ends at 83.75 (72.5+11.25]. Finally the fourth interval is (83.75, 95.0]. All the intervals along with the corresponding temperature attribute values are shown below:

(50.0-ε, 61.25] → [50]
(61.25,72.5] → [64,65,68,69,70,71,72]
(72.5,83.75] → [73,75,75,80,81,83]
(83.75,95.0] → [85,95]

b) In Binning by Equal-depth, the number of items in each interval are the same. Number of items per interval is calculated by dividing the total number of items by the number of bins. In our case calculation is done on the HUMIDITY attribute. Total number of items is 16 and the number of bins is 4, therefore the number of items per each bin is 4. The HUMIDITY attribute items are listed below:

HUMIDITY = [85,90,86,96,80,70,65,95,70,80,71,89,75,91,85,45]. For getting intervals sort the data , we get HUMIDITY = [45, 65, 70, 70, 71, 75, 80, 80, 85, 85, 86, 89, 90, 91, 95, 96]. Therefore the four intervals are:

1 -> [45,65,70,70]
2 -> [71,75,80,80]
3 -> [85,85,86,89]
4 -> [90,91,95,96]

c) In this case we discretize in intervals of $[x_c +(k-1)\sigma, x_c + k \sigma)$. Here σ is standard deviation which is given as 13 and x_c is the mean which is given as 80. The intervals have to be calculated until all the items are covered in the intervals.

For k = 1, we get [80, 80+13) -> [80,93)
For k = 2, we get [93, 80+2*13) -> [93,106)

From the above two intervals all the values greater than the mean are covered from the HUMIDITY items. For the items less than the mean we take value of k =0,-1,-2,…..

For k= 0, we get [80-13,80) ->[67,80)

For k = -1, we get [80-2*13,67) -> [54,67)

For k = -2, we get [80-3*13,54) -> [41,54)

The final intervals along with the corresponding items are:

[41,54) -> [45]

[54,67) -> [65]

[67,80) -> [70,70,71,75]

[80,93) -> [80,80,85,85,86,89,90,91]

[93,106) -> [95,96]

d)

i) Equal-width binning:

Advantages: It is a straightforward method of dividing the data into intervals where each interval has items of a specific range.

Disadvantages: But there might be problems if there are outliers in the data which might disrupt the intervals. Skewed data is not handled properly in this method.

ii) Equal-depth binning:

Advantages: It is useful for good scaling of data. This method is used when we require intervals with equal number of items in them.

Disadvantages: This method does not handle the categorical attribute data properly. It is difficult to calculate the number of bins.

iii) Third method:

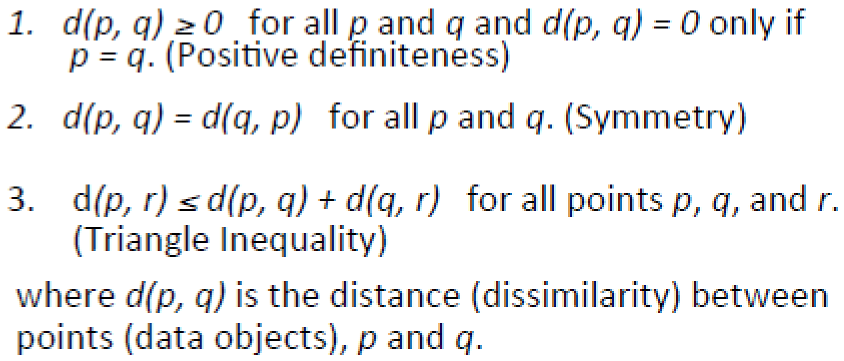
Advantages: This method helps in finding out how far each item is from the mean of the data. The intervals have a width of σ which is the standard deviation of the data. There is no constraint on the number of bins to be used.

Disadvantages: There is a possibility that the data may not be properly distributed across the bins. For example if the data is present close to mean as seen in the case of question 5c.

([80,93) -> [80,80,85,85,86,89,90,91]). This method is vulnerable to outliers.

6) Distance Metrics

a) As per the notes, below are the definitions for positive definiteness, symmetry and triangle inequality.



1. *d(p, q)* $\geq$ *0* for all *p* and *q* and *d(p, q) = 0* only if *p = q*. (Positive definiteness)
2. *d(p, q) = d(q, p)* for all *p* and *q*. (Symmetry)
3. d*(p, r)* $\leq$ *d(p, q) + d(q, r)* for all points *p*, *q*, and *r*. (Triangle Inequality)

where *d(p, q)* is the distance (dissimilarity) between points (data objects), *p* and *q*.

i) Euclidean distance formula is $((x2-x1)^2 + (y2-y1)^2)^{½}$. This satisfies the positive definiteness property. The value $(x2-x1)^2 + (y2-y1)^2$ will always be positive since we are adding squares of two numbers and a square of a number is always positive. When we take the square root of this positive value, and the square root will also be positive. Also if the points are the same , then x2=x1 and y2=y1. Hence we get x2-x1=y2-y1 = 0, square of 0 is 0 and square root is also 0. Hence the d(p,q) = 0 only when p=q. The symmetry property is satisfied since we are squaring the difference between coordinates and the squares of a negative or positive number with the same absolute value is the same. Hence d(p,q) = d(q,p). The triangle inequality property is also satisfied since the straight line euclidean between any two points will always be lesser than going through a triangular intermediary point.

ii) Manhattan distance formula is |x2-x1|+|y2-y1|. This satisfies the positive definiteness property as the absolute value of the difference between the coordinates will always be a positive value too. Also, the difference between both the x and y coordinates in this case will be 0 if and only if it is the same point. Then we will be calculating |0|+|0|. This distance metric also satisfies symmetry, since we are taking the absolute value it doesn't matter if we swap x2 with x1 and x1 with x2 while taking the difference. Same goes for the y coordinates. Hence it doesn't matter if we are calculating d(p,q) or d(q,p) since both will yield the same value due to the absolute factor in the manhattan distance. This also satisfies the triangle inequality property as d(p,q) <= d(p,r) + d(r,q). This is because a triangle is made up of straight lines and hence the offset needed by d(p,q) which arises due to the manhattan distance formula is similarly compensated for in the d(r,q) with its own offset to get to q.

iii) Divergence function formula is d(A,B) = 1 - |A and B|/|A|. This function doesn't satisfy the positive definiteness property. Although |A and B| is always |A and B| <= |A|, but If |A and B| = |A| then that means B = A or B is a superset of A.

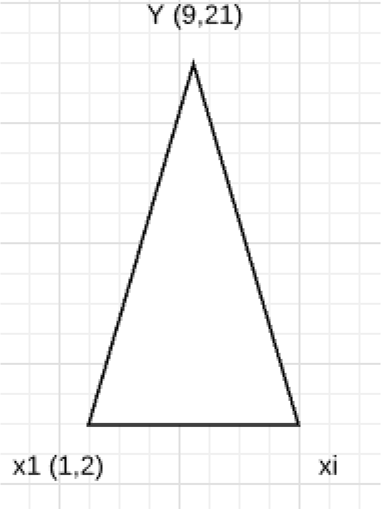
This means d(A,B) = 0 even if A not equal to B and both are not the same sets, but instead B is a superset of A containing all the elements of A. Eg. Lets say A has 10 items, and B has 10+10 items of A = total 20 items. Then d(A,B) = 1-10/10 = 0, since A and B will be equal to |A|. This function also does not satisfy the symmetry property as the denominator changes in |A and B|/|A| depending on what A and B is. Eg. consider A has 10 items and B has 20 items with 5 being common between the two. Then d(A,B) = 1-5/10 = 0.5 but d(B,A) = 1-5/20 = 0.75. Hence d(A,B) not equal to d(B,A). This function also does not satisfy the triangle inequality property. Eg. Let's consider 3 sets A, B, C where A ={1,2,3,10}, B={1,2,4,5,6,7,8,9} and C={1,4,5,6,7,8,9}. Now d(A,C) = 1-¼ = ¾. d(A,B) = 1-½ =½. d(B,C) = 1-⅞ =⅛. d(A,B) + d(B,C) = ⅝ . But we have d(A,C) = ¾ which is not less than or equal to sum of d(A,B) and d(B,C). Hence the property d(A,C) <= d(A,B) + d(B,C) is not satisfied. Hence each computation is not mutually exclusive, due to which this property is violated.

iv) The formula for cosine distance is d(A,B) = 1 - A.B/(||A|| ||B||). This does not satisfy the positive definiteness property. Even though the cosine similarity always lies between 0-1 and d(A,B) = 0 if A= B, this is not the only case when d(A,B) is 0. For example let's take 2 points on a line which makes an angle of 45 degrees with the X axis. Let's say these points are (3,3) and (4,4). In this case also d(A,B) = 0 but A is not equal to B. Hence the positive definiteness property is not satisfied. Cosine distance satisfies the symmetry property. This is because dot product of any two vectors is commutative. That means A.B=B.A. Hence if we run the formula, this translates to d(A,B)=d(B,A). The triangle inequality is not satisfied in cosine distance function. This is primarily because cosine similarity function which is used internally does not satisfy the triangle inequality property. Eg. Let's say there are 3 unit vectors A=(1,0), B=($2^{1/2}$/2,$2^{1/2}$/2), C=(0,1). Hence, d(A,C) = 1-0 = 1 . d(A,B) = 1- $2^{1/2}$/2 = 0.293 . d(B,C) = 1- $2^{1/2}$/2 = 0.293. Hence 1 > 0.293+0.293 which violates the triangle inequality.

b)

i) The triangle inequality property will help us reduce the number of comparisons in the 1 NN algorithm. To do this we basically shuffle the terms around in the triangle property formula and make use of the fact d(a,b) >= |d(a,c) - d(c,b)|

ii) We can make use of the triangle inequality property to skip certain points in the training data set, so that they are never compared with Y and the cost of d(y,x_i) is saved. Consider the below triangle, xi represents points in the training set. Let's say the points in the training set are {(1,2),(2,3),(4,5),(8,20)}. Let's say our y is (9,21). Let's pick a random point from the training set as our base point, using which we will be forming the below triangle with all other points in the training set. Let's say the point we pick is x1 (1,2). Distance between x1 and y is d(x1,y) = 20.61. Now the triangle property is basically d(xi,x1) <= d(x1,y) + d(xi,y). This translates to d(xi,y) >= |d(xi,x1) - d(x1,y)|. We take the absolute value in this case as distance cannot be negative. Now, using this formula, we calculate the minimum value any d(xi,y) can take which is equal to |d(xi,x1) - d(x1,y)|.The actual value of d(xi,y) can either be greater than or equal to this value. The corresponding minimum values for all d(xi,y), where i goes from 1 to 4 for all x in the training set are {20.61,19.2,16.37,1.3}. Now looking at this, we can note that our best guess to try will be the point which has the minimum value in this set. Hence, let’s rearrange the set in ascending order => {1.3,16.37,19.2,20.61}. The value 1.3 corresponds to the point (8,20), lets note this and calculate the actual distance between this point and y. This comes out to be 1.414. Now we go to the next point that is (4,5) with value 16.37. But as d(x3,y) can be minimum 16.37, there is no way it can be equal or lesser than the distance 1.414 which we have already got. Hence we can terminate the algorithm and return the closest point to y as x4 (8,20). In this case we have just compared y with the base point x1 and x4, skipping comparisons with all other points in the process. This was done using the property of triangle inequality.

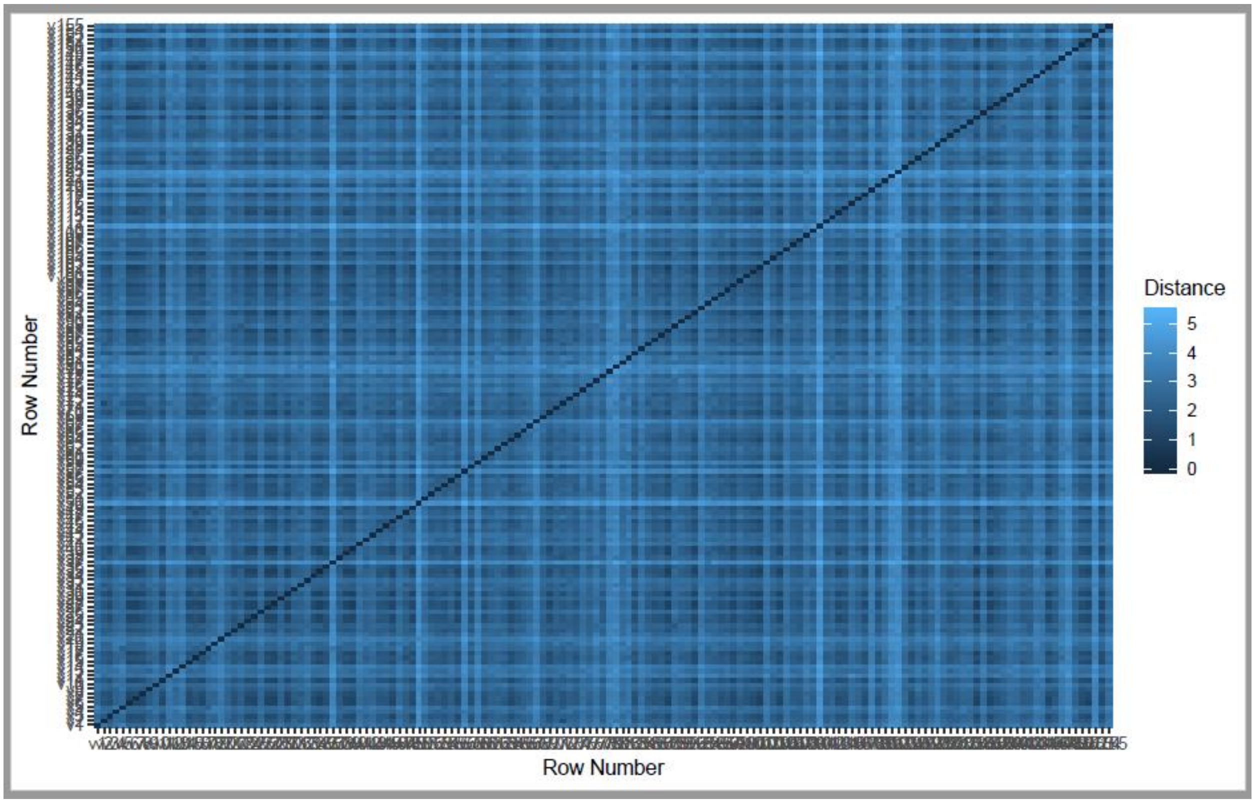


iii) The best case over here will be when the data points in the training sets are scattered away from each other. Basically this will give us a situation where the test data point, will get the closest point in the first go itself as done above, without having to go any further. So, in this case we will just need to do two comparisons with y. First is the comparison of y with the base point x1. The second comparison is to calculate the actual value between y and the nearest candidate let’s say this distance is d1. As the other points will be very far off, the minimum distance between those points and y will be very high when compared to d1. Hence we terminate our algorithm and return the nearest point in X to y.

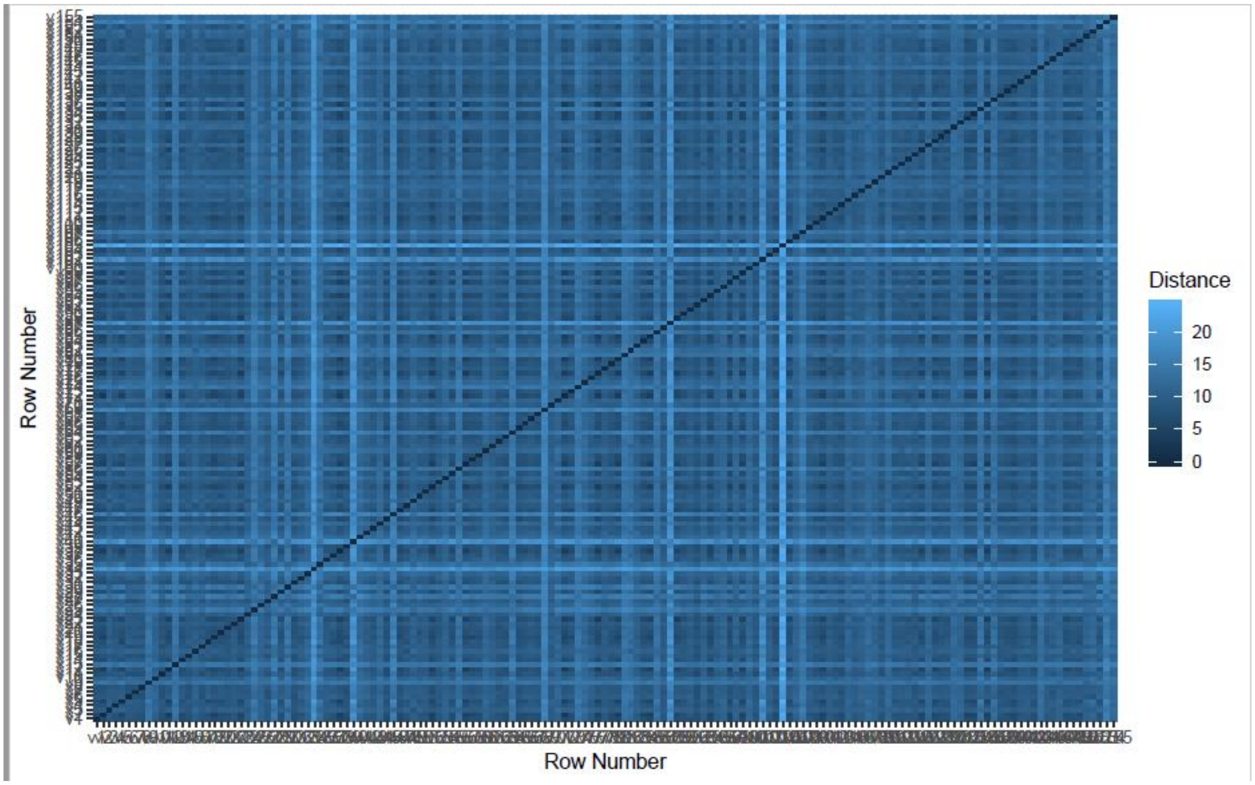
The worst case will be when the training data points are very close together. Then in this case the distance of any point xi from y will be almost similar. Hence we will end up going through all the points, finally checking the actual distances between xi and y before figuring out the nearest neighbour of y in the training data.

7) Original matrix <- Euclidean over original data matrix read from csv file.
Normalized matrix <- Euclidean over Normalized data matrix.
Above terms used in the explanation below.

Normalized Matrix



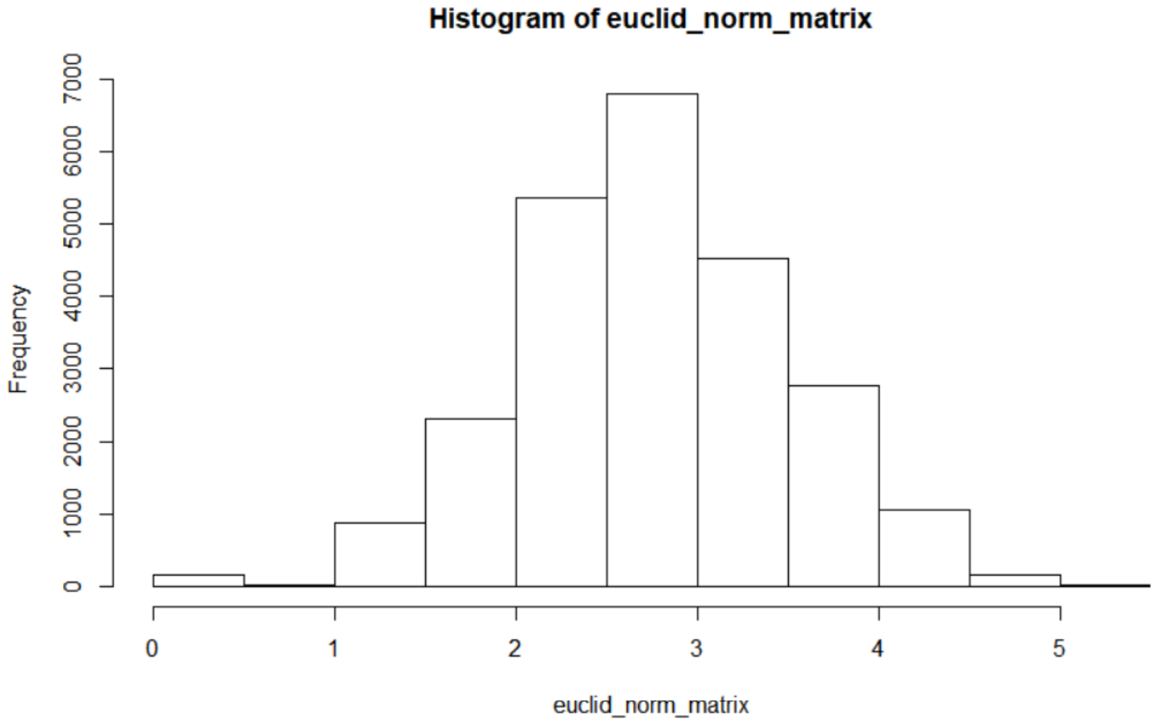
Original Matrix

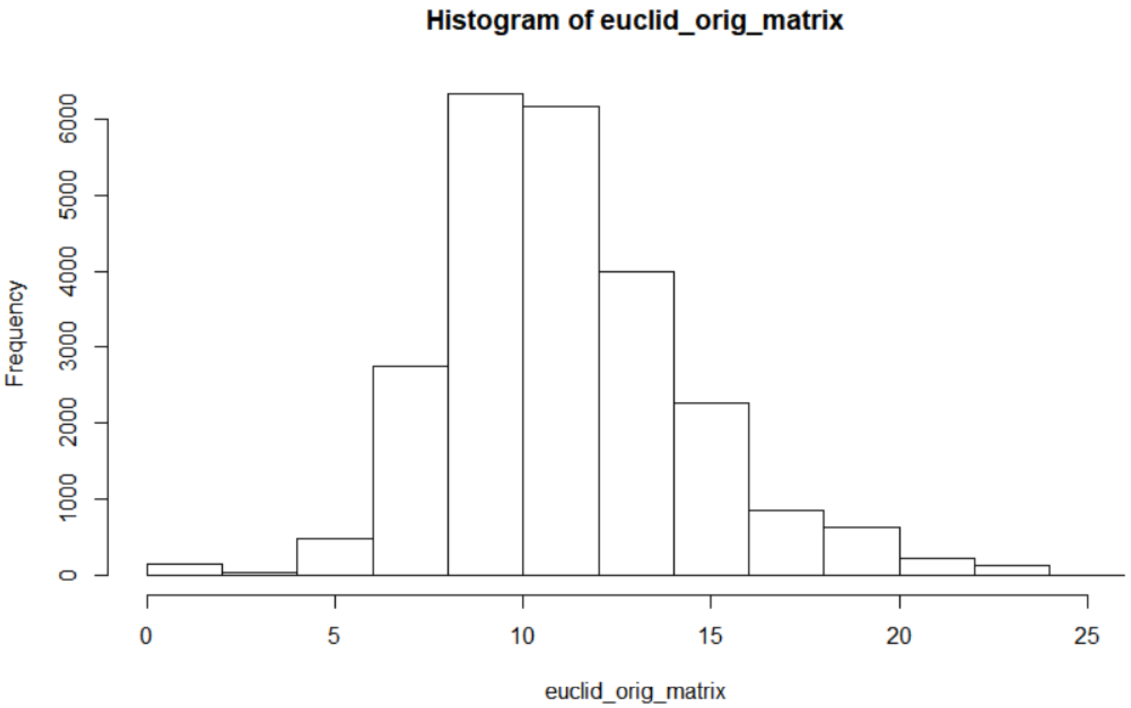


i) By observing the plots for the two matrices, we see that the maximum distance in Normalized matrix smaller when compared to the Original data matrix. The maximum distance for Normalized matrix is ~5 and Original matrix is ~24. This change is observed as the Normalized matrix scales the original matrix values to [0,1].

ii) The mean and standard deviation for the two matrices vary. The observed mean value for Normalized matrix is 2.75 and original matrix is 11.13. Similarly the observed standard deviation values for Normalized and Original matrices are 0.74 and 3.31 respectively. In Normalized matrix all the points contribute proportionately for calculation of mean or standard deviation. The reason for this being it is scaled to certain range. In the case of our original matrix there might be certain values which might contribute more.

iii) If we look below into the histograms of two matrices we see the distribution of data in the original matrix to be partially skewed. The histogram of Normalized matrix shows the data is distributed with no skewness. This is due to the reason as mentioned above that is the data points in Normalized matrix contribute proportionately for any measure.


Histogram of euclid_norm_matrix
Frequency
0
1000
2000
3000
4000
5000
6000
7000
0
1
2
3
4
5
euclid_norm_matrix


Histogram of euclid_orig_matrix
Frequency
0
1000
2000
3000
4000
5000
6000
0
5
10
15
20
25
euclid_orig_matrix